EIEF

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF WORKING PAPER SERIES

EIEF Working Paper 20/13

May 2020

Estimating the prevalence of the COVID-19 infection, with an application to Italy

by
Franco Peracchi

(Georgetown University, EIEF, and University of Rome Tor Vergata)

Daniele Terlizzese

(EIEF)

Estimating the prevalence of the COVID-19 infection, with an application to Italy[*]

Franco Peracchi
Georgetown University, EIEF, and University of Rome Tor Vergata

Daniele Terlizzese
EIEF and Banca d'Italia

August 4, 2020

Abstract

Knowing the prevalence of the COVID-19 infection in a population of interest, and how it changes over time and across space, is of fundamental importance for public health. Unfortunately, the fraction of cases who turn out to be positive in a test provides a distorted picture of the prevalence of the infection because the tested cases are not a random sample of the population. Since random testing of the population is costly and complicated to carry out, in this note we show how to use the available information, in conjunction with credible assumptions about unknown quantities, to obtain a range of plausible values for the prevalence of the infection. We discuss the difference between two alternative measures of prevalence and we argue that one of the two is much harder to pin down with the data currently available. We apply our method to the Italian data.

Keywords: COVID-19 infection; point prevalence; period prevalence; Italy.

JEL codes: C29, C83, I19

[*] Corresponding author: Daniele Terlizzese (daniele.terlizzese@eief.it). We thank Giulia Faggio, Luigi Guiso, Chuck Manski, Francesca Molinari, and Antonia Stazi for helpful comments. We also thank the press office of Dipartimento della Protezione Civile and officers at the Italian Ministry of Health for detailed information about the COVID-19 data.

1 Introduction

Knowing the prevalence of the COVID-19 infection in a population of interest, and how it changes over time and across space, is of fundamental importance for public health. Unfortunately, the fraction of people who are confirmed positive to COVID-19, reported daily in different countries by the agencies monitoring the evolution of the pandemics, provides a biased picture of the prevalence of the infection, since people tested for the presence of the virus are not a random sample of the population. In fact, tests are mainly done on subjects with severe enough symptoms or high likelihood of a contact with infected cases, and the pool of those tested grossly underrepresents the asymptomatic and those with only mild symptoms. As the latter are a powerful vehicle of contagion, the lack of information on their prevalence is a serious hindrance to the design of effective containment policies.

This is of course well known (see, e.g., Li et al., 2020 and Stock, 2020), and it is understandable that, in the heat of the crisis, attention is focused on saving lives rather than having a statistically accurate representation of the phenomenon. However, the correct measurement of the prevalence of the infection is starting to be perceived as an essential input in planning a gradual removal of lockdown policies. For example, the Italian national statistical agency (Istituto Nazionale di Statistica, or Istat) just launched a relatively large survey (with a target sample size of 150 thousands individuals) to ascertain, through serological testing, the proportion of the population that has been infected by the virus. But these surveys are costly, complicated to carry out, take time to complete and may be plagued by (non random) refusal to participate. It is therefore of some value to come up with easy to implement, though necessarily less accurate, estimates of the prevalence of the infection in the population based on readily available data.

Indeed, there are two somewhat different concepts of prevalence. One is point prevalence, namely the fraction of people in the population who, *at a given point in time*, are infected; they might have contracted the infection in a previous date, but are still infected at the point in time we are considering. The other is period prevalence, namely the fraction of people in the population who have been infected during a given period; in particular, if the period is from the beginning of the pandemic to the current moment, period prevalence measures the fraction of people in the population who were *ever* infected, even though at the present date might no longer be – because recovered or dead. Point prevalence gives information on the evolution of the contagion since it measures the fraction of people who are still infectious. It is therefore more useful to monitor the risk of transmitting the infection. Period prevalence is instead more useful to monitor the risk of being infected – since it readily gives the fraction of people susceptible of being infected – and for reporting purposes, for example to correctly compute the case fatality ratio or to assess the productivity of the resources deployed to fight the infection. Depending on the nature of the tests conducted, different surveys would measure one or the other concept of prevalence: roughly speaking, a survey based on nasopharyngeal swabs, tested for the presence of the virus, aims at measuring point prevalence, while a survey based on serological tests, that detect the antibodies produced in response to the infection, aims at measuring period prevalence.

In this short paper we present an approach to obtain estimates of point prevalence based on readily available data and minimal, transparent, and easy to interpret assumptions. A recent paper by Manski and Molinari (2020) has a similar goal and approach but focuses on period prevalence. Since both concepts of prevalence are interesting and useful, the two papers nicely complement each other. We then apply our approach to data from Italy, both for the country as a whole and for its different regions.

Estimates of the point or period prevalence have also been produced in the context of structural or semi-structural models of the epidemic (see for example Flaxman et al., 2020 and Vollmer et al., 2020). The main advantage of these models is their stylized description of some of the mechanisms underlying the epidemic. However, the resulting estimates typically require strong and untested parametric assumptions, and the complexity of these models makes the mapping between the assumptions and the estimates more opaque.

The paper proceeds as follows. Section 2 outlines our approach. Section 3 offers a brief description of the data we use. Section 4 presents our empirical results. Section 5 clarifies in a simplified setting the links between point and period prevalence, and Section 6 briefly concludes.

2 Framework

To avoid complicating the analysis, we assume that each person can have at most one episode of infection (after which she either recovers or dies), that is infectious as long as she is infected, and that the recovered are immune and can no longer transmit the infection. These assumptions, though not strictly true, are close enough to what is known about the COVID-19 pandemic. For the sake of simplicity, we also neglect the effect of deaths and new births on the population size, which we take as constant. Time is measured in days.

2.1 Notation

Let I_t be a binary random variable that is equal to one if a person is infected on day t, and is equal to zero otherwise. An infected person on day t may have become infected before t, and some of those who became infected before t may have recovered (or died) by day t; for them $I_t = 0$ would hold. The quantity that we want to estimate is $\Pr(I_t = 1)$, which we interpret as the population fraction of individuals who on day t are infected, and therefore can infect others.

Also, let T_t be a binary random variable that is equal to one if a person has been first tested on day t for the presence of the virus through a nasopharyngeal swab, and is equal to zero otherwise. Notice that, if $T_t = 1$, then $T_{t+k} = 0$ for all $k \geq 1$. Further, $T_t = 0$ if either a person was first tested at some previous date or was never tested. We interpret $\Pr(T_t = 1)$ as the population fraction of individuals who were first tested on day t.

Finally, let P_t be a binary random variable that is equal to one if a person has first received a positive test result on day t, and is equal to zero otherwise. Hence, if $P_t = 1$, then $P_{t+k} = 0$ for all $k \geq 1$. Further, $P_t = 0$ if either a person was first tested on day t with a negative test result, or was not first tested on that day, i.e., was first tested on some previous day or was never tested.

It is convenient to assume that people are tested at most once, and that test results are available the same day when the swabs are taken. If follows from these two assumptions that $P_t = 1$ implies $T_t = 1$, that is, those who first receive a positive test result on day t must have been first tested on that day.[1] Although neither assumption is exactly true, we will discuss in Section (3) how to transform the data to get a reasonable approximation.

2.2 The estimating equation

To simplify the notation, in this section we drop the time subscript t. By the Law of Total Probability, we can write

$$\Pr(I = 1) = \Pr(I = 1|T = 1)\ \Pr(T = 1) + \Pr(I = 1|T = 0)(1 - \Pr(T = 1)). \tag{1}$$

Since $\Pr(T = 1)$ is directly measured in the data, to estimate $\Pr(I = 1)$ we need information about $\Pr(I = 1|T = 1)$ and $\Pr(I = 1|T = 0)$. Let us consider these two probabilities separately.

As for $\Pr(I = 1|T = 1)$, we can exploit the available information on the operational properties of the tests conducted on nasopharyngeal swabs, namely their Type-I and Type-II error probabilities. There seems to be widespread consensus that these tests have a probability of Type-I error – the probability of a false positive – that is very close to zero.[2] We therefore assume that

$$\Pr(P = 1|I = 0, T = 1) = 0. \tag{2}$$

This assumption has two important implications. The first is that

$$\Pr(I = 1|P = 1) = 1, \tag{3}$$

[1] The assumption that nobody is tested more than once rules out the case of a person who is first tested on day t, resulting negative, and then retested at a later date, resulting positive for the first time. This case would contradict the implication claimed in the text.

[2] In the Italian case, for a period at the beginning of the pandemic, all swabs found positive by regional labs were retested by the Istituto Superiore di Sanità (the Italian equivalent of the National Institute of Health) and always confirmed. This reinforces the assumption of zero probability of a false positive.

which follows directly from Bayes Law and the already established implication that $T = 1$ if $P = 1$. The other is that

$$\Pr(I = 1|T = 1) = \frac{\Pr(P = 1|T = 1)}{1 - \beta}, \tag{4}$$

where $\beta = \Pr(P = 0|I = 1, T = 1)$ is the probability of Type-II error – the probability of a false negative. To see this, notice that

$$\begin{aligned}
\frac{\Pr(P = 1|T = 1)}{1 - \beta} &= \frac{\Pr(P = 1|T = 1)}{\Pr(P = 1|I = 1, T = 1)} \\
&= \frac{\Pr(P = 1, T = 1)}{\Pr(T = 1)} \times \frac{\Pr(I = 1, T = 1)}{\Pr(I = 1, T = 1, P = 1)} \\
&= \frac{\Pr(P = 1)}{\Pr(T = 1)} \times \frac{\Pr(I = 1, T = 1)}{\Pr(I = 1, P = 1)} \\
&= \frac{\Pr(I = 1|T = 1)}{\Pr(I = 1|P = 1)} \\
&= \Pr(I = 1|T = 1),
\end{aligned}$$

where the first equality uses the definition of β, the second and the fourth follow from the rules of conditional probabilities, the third from the fact that $T = 1$ if $P = 1$, which implies that $\Pr(P = 1, T = 1) = \Pr(P = 1)$ and $\Pr(I = 1, T = 1, P = 1) = \Pr(I = 1, P = 1)$, and the last from (3).

Equation (4) is important because it expresses $\Pr(I = 1|T = 1)$ as a function of quantities for which we either have direct measures, namely $\Pr(P = 1|T = 1)$, or we can make an educated guess based on medical knowledge, namely β. Although estimating the probability of a false negative is not easy, there is a general consensus in the medical profession that it is not negligible and largely reflects issues with specimen collection; specifically, the sample might have been collected too early or too late, might be contaminated, or might have been stored for too long.[3] A review of the available health literature suggests a range of values for β between .02 and .4, though a narrower range between .1 and .3 is more often quoted.[4] To the extent that β reflects practical issues in the implementation of the test, one might expect some time variability due to learning. However, there is little evidence of this and we will assume that β is constant.

As for $\Pr(I = 1|T = 0)$, we rely on information about the testing process and assume that

$$\Pr(I = 1|T = 0) \leq \Pr(I = 1|T = 1). \tag{5}$$

This is because the subjects with $T = 0$ on day t are those who either have not been tested yet or have been tested at an earlier date. Because nasopharyngeal swabs are mainly taken from subjects who have visible symptoms, or are suspected of exposure to the infection, those who have not been

[3] https://asm.org/Articles/2020/April/False-Negatives-and-Reinfections-the-Challenges-of.

[4] See for example Watson et al., 2020. Also see https://www.medpagetoday.com/infectiousdisease/covid19/86047, https://www.healthline.com/health-news/false-negatives-covid19-tests-symptoms-assume-you-have-illness, and https://theconversation.com/coronavirus-tests-are-pretty-accurate-but-far-from-perfect-136671.

tested are clearly less likely to be infected as of day t; moreover, among those who have been tested at an earlier date, some will have recovered and are no longer infected. Therefore, it seems reasonable to assume that subjects with $T_t = 0$ are on average less likely to be infected than those with $T_t = 1$.[5] Thus, we define

$$\lambda = \frac{\Pr(I = 1|T = 0)}{\Pr(I = 1|T = 1)},$$

and assume that λ ranges between 0 and 1. If people are tested at random, the infection rate is the same among the tested and the untested, and therefore $\lambda = 1$. If the tested sample is instead biased towards the symptomatic (or, more generally, towards those with higher infection risk), then $\lambda < 1$.

Putting it all together we can rewrite (1) as

$$\Pr(I = 1) = \frac{\Pr(P = 1|T = 1)}{1 - \beta} \left[\Pr(T = 1) + \lambda \left(1 - \Pr(T = 1)\right)\right]. \tag{6}$$

This equation is our basis for estimating $\Pr(I = 1)$ as a function of λ and β given knowledge of two other quantities that we can measure in the data, namely $\Pr(P = 1|T = 1)$ and $\Pr(T = 1)$.

2.3 The anatomy of λ

As already mentioned, we can bring information obtained from medical expertise and testing practice to bear on β. The natural question is whether something similar can be done about λ. To do this, it is useful to derive an expression for λ in terms of quantities that are potentially observable.

We distinguish between symptomatic ($S = 1$) and asymptomatic ($S = 0$) cases, where the former are those who show some of the symptoms associated with the infection and the latter show none of them, and let $\Pr(I = 1|S = s)$, $s = 0, 1$, be the fraction of the infected among the two types. It is reasonable to assume that the infection rate is higher among the symptomatic, so we posit

$$\mu = \frac{\Pr(I = 1|S = 1)}{\Pr(I = 1|S = 0)} \geq 1.$$

Let $n = \Pr(T = 1)$ be the fraction of the population that is tested, $\pi = \Pr(S = 1)$ the fraction of the population that is symptomatic, and $p = \Pr(S = 1|T = 1)$ the fraction of the symptomatic among the tested. Therefore $\gamma = p/\pi$ is a measure of the bias implicit in the testing protocol, which we know targets disproportionally symptomatic subjects. This implies that $\gamma \geq 1$.

By the Law of Total Probability, we can write the infection rate among the tested cases as

$$\Pr(I = 1|T = 1) = \Pr(I = 1|S = 1, T = 1)\, p + \Pr(I = 1|S = 0, T = 1)(1 - p).$$

We know that the bias in the testing protocol is mainly due to the higher proportion of symptomatic case been tested; to obtain a sharp characterization, we then make the extreme assumption that the

[5] Assumption (5) is similar to the monotonic testing assumption in Manski and Molinari (2020), though we refer here to point prevalence while they refer to period prevalence.

bias is entirely due to the oversampling of the symptomatic subjects. Specifically, we assume that, conditional on symptomatology, the infection rate is the same in the tested sample and the general population.[6] Formally, we assume

$$\Pr(I = 1|S = s, T = 1) = \Pr(I = 1|S = s), \quad s = 0, 1. \tag{7}$$

Under this assumption

$$\begin{aligned} \Pr(I = 1|T = 1) &= \Pr(I = 1|S = 1)\, p + \Pr(I = 1|S = 0)(1 - p) \\ &= [(\mu - 1)p + 1]\ \Pr(I = 1|S = 0), \end{aligned}$$

where we used the definition of μ. Among the untested, the infection rate is instead

$$\begin{aligned} \Pr(I = 1|T = 0) &= \frac{\Pr(I = 1) - \Pr(I = 1, T = 1)}{\Pr(T = 0)} \\ &= \frac{\Pr(I = 1|S = 1)\, \pi + \Pr(I = 1|S = 0)(1 - \pi) - \Pr(I = 1|T = 1)\ \Pr(T = 1)}{\Pr(T = 0)} \\ &= \frac{(\mu - 1)(\pi - np) + 1 - n}{1 - n}\ \Pr(I = 1|S = 0). \end{aligned}$$

Combining these two results, using $p = \gamma\pi$, we obtain

$$\lambda = \frac{(\mu - 1)(1 - n\gamma)\pi + 1 - n}{(1 - n)[\gamma\pi(\mu - 1) + 1]}. \tag{8}$$

This equation shows that λ can be represented in terms of four quantities – the share of the population tested (n), the share of the symptomatic in the population (π), the ratio of the infection rates among the symptomatic and the asymptomatic (μ), and the bias of the tested sample towards the symptomatic (γ) – on some of which health authorities may have updated and detailed information.[7] This knowledge could in turn be exploited to narrow the range of possible values for λ and therefore, through (6), the range of possible values of $\Pr(I = 1)$.

We illustrate by showing how the results of the near-complete testing conducted in a small town in Northern Italy can be used to this effect. Specifically, we take advantage of the data from Vo', the municipality of the Veneto region with the first COVID-19 death in Italy (on February 21, 2020), that conducted (twice) an almost complete testing of its population.[8] These data allow us to compute the population fraction of symptomatic (people with fever or cough or other symptoms like respiratory

[6] While it is likely true that, even among the symptomatic cases, the doctors administering the test might further discriminate, targeting those with more severe or more numerous symptoms, we believe that our assumption captures the bulk of the bias.

[7] Note that μ and γ both measure relative differences: how much more infected are the symptomatic vs. the asymptomatic, and how much higher is the fraction of symptomatic in the tested sample than in the general population. As such, they are arguably easier to assess, and possibly to calibrate with available evidence.

[8] The data from Vo' contain the results of two separate surveys carried out between February 21 and March 10, 2020 (covering 2,812 and 2,343 subjects respectively, corresponding to 86% and 72% of the eligible population of the municipality), and have been downloaded from the supplementary material of Lavezzo et al. (2020).

difficulties or diarrhoea) and asymptomatic cases (people without any of the symptoms related to COVID-19) – these were .067 and .933, respectively; we then calibrate π at .067. For n we take the value .0004 (the testing rate in Italy on June 20, 2020; see Section 3). Given these two values, from equation (8) we can express λ as a function of γ and μ.

Figure 1 shows the contour plot of this function, for $1 \leq \gamma \leq 1/\pi$ (which ensures that $p \leq 1$) and $1 \leq \mu \leq 25$. The color intensity ranges from dark green in the lower left corner, for values of λ close to one (the value we would observe with random sampling) arising from a combination of values of γ and μ close to one, to light green in the upper right corner, for values of λ close to zero arising from a combination of high values of γ and μ.

In the case of Vo', where the fraction of positive cases is .232 among the symptomatic and .0129 among the asymptomatic, we observe $\mu = 17.9$. For this value of μ, λ would vary between 0.11, when the fraction of symptomatic in the sample is 15 times the fraction in the population, and .65, when the sample contains twice as much symptomatic subjects as in the population (i.e., $\gamma = 2$), a value that we take as the lower bound of the bias intensity.

We do not know whether the value of μ observed in the data from Vo' between late February and early March 2020 is representative of the relative prevalence of COVID-19 among symptomatic and asymptomatic subjects in different regions of Italy or at different times into the evolution of the pandemic. Nevertheless, we think it provides a useful benchmark. More importantly, we think that the health authorities might have a better sense of the size of the bias implicit in the testing protocol, and equation (8) might then be used to narrow down the values for λ.

3 Data

We confine attention to Italy, breaking down the analysis to the regional level to account for the large geographical heterogeneity. We use two data sources: Istat for the Italian population at the onset of COVID-19 and the Italian Department of Civil Protection (Dipartimento della Protezione Civile, or DPC) for daily summaries of the epidemic.

The Istat data are estimates of the resident population as of January 1, 2019, broken down by region (Istituto Nazionale di Statistica, 2019).

The DPC data are the product of the data collection effort coordinated by the Italian integrated COVID-19 surveillance system (Riccardo et al., 2020) and consist of daily time-series at various levels of geographic aggregation. For our purposes, the most important series are: the total number of detected cases ("totale casi"), the number of new detected cases ("nuovi positivi"), equal to the daily change in the total number of detected cases, the total number of swabs ("tamponi"), and the total number of tested cases ("casi testati"). The first three series are available at the national and regional level from February 24, 2020, the last only from April 19.

A number of remarks on the DPC series are in order. First, the series of new detected cases gives the number of subjects who *first* test positive on day t – i.e., the number of subjects for whom $P_t = 1$ – irrespective of whether they became infected on day t or earlier.

Secondly, the number of swabs overstates the number of people actually tested because certain subjects are tested repeatedly. In addition to health personnel and other people employed in critical services, these include people discharged from a hospital who need to test negative at least twice before being sent back home. Besides duplications, the timing between swabs and tests is not fully aligned. In particular, the change in the number of swabs between day $t-1$ and day t actually records the number of tests results obtained on day t. Due to delays in processing and reporting, these include test results from swabs taken before t and excludes swabs taken in t but not yet processed, or for which test results are not yet available.

Thirdly, the series of tested cases records the total number of subjects from whom a swab was taken, thus eliminating the duplications contained in the swabs series. To estimate $\Pr(P = 1|T = 1)$, it is therefore more appropriate to take the ratio between detected and tested cases, rather than the ratio between detected and swabs. We think that the advantage of excluding the duplications outweighs some drawbacks associated with the series of tested cases, namely the shorter time span and two issues that arise because of the delays in test processing and the retesting rules. The first issue is timing misalignments between swabs and tests, which we just discussed. We address this issue by taking centered moving averages of daily changes of both detected and tested cases. The second issue instead arises because the series of new tested cases excludes subjects who, after being tested negative, are retested at a later date. This, although consistent with our assumption that no subject is tested more than once, leads to overstate the prevalence of the infection among the tested. We have no fix for this issue, but we conjecture that, given the tight constraints on testing capacity, the number of subjects retested after a negative result is relatively small, except possibly for some health personnel and workers in other critical services.

A final warning concerning the data is that the series of tested cases in one of the regions (Lazio) has a clear break on April 24, most likely because of an initial reporting error. We will therefore consider the data starting from April 25.

4 Results

To produce estimates of $\Pr(I = 1)$ based on equation (6), we need to compute the population fraction of new tested cases, $\Pr(T = 1)$, and the population fraction of new positive among the new tested cases, $\Pr(P = 1|T = 1)$. We compute these two fractions using the population estimates from Istat and the DPC data starting from April 25, 2020. For the reason explained in Section 3, we present results for a 7-day centered moving average, though results for 3-day and 5-day moving averages are very similar. A 7-day moving average considerably reduces the sample period but has the advantage of removing the day-of-week effects that are clearly present in the data.

Table 1 presents the observed values of $\Pr(T = 1)$ and $\Pr(P = 1|T = 1)$ as of June 20, 2020. Only a small fraction of the population is newly tested in a given day (.04 percent on average), with substantial variation across regions, from .01 percent in Campania to .08 percent in the Northern regions of Emilia-Romagna, Friuli Venezia Giulia and Trentino-Alto Adige. The fraction of new positives among the newly tested is .78 percent on average, with much larger variation across regions, from less than .005 percent in Basilicata to 2.7 percent in Lombardia, reflecting substantial regional differences in both the intensity of the epidemic and the bias in testing. The correlation between $\Pr(P = 1|T = 1)$ and $\Pr(T = 1)$ is positive but very weak (less than .10).

To produce estimates of $\Pr(I = 1)$, we also need to assign values to β and λ.[9] For the former, as already mentioned, the available information suggests a relatively narrow range between .10 and .30, which we conservatively broaden to the range between .01 and .50. For the latter, we have little a priori information. We showed in Section 2.3 that, using data from a small town in the Veneto region, λ could plausibly be narrowed down to the range $[.10, .65]$. Again conservatively, we report results for the wider range $[.01, .99]$.

Table 2 presents our estimates of $\Pr(I = 1)$ for Italy as a whole, as of June 20, 2020. The values in red correspond to $.10 \leq \lambda \leq .65$ and $.10 \leq \beta \leq .30$ and range between .10 percent when $(\lambda, \beta) = (.10, .10)$ and .70 percent when $(\lambda, \beta) = (.65, .30)$. Given the Italian population of 60.4 million, this corresponds to a range between 60 and 423 thousands infected people as of June 20, 2020.

Table 3 shows our estimates, for Italy as a whole and separately by region, for three illustrative pairs of values of (λ, β), namely $(.25, .10)$, $(.50, .20)$, and $(.75, .30)$. Figure 2 instead shows the value of $\Pr(I = 1)$ by region at a point in time, namely June 20, 2020, for $(\lambda, \beta) = (.50, .20)$. The estimates reveal a clear North-South gradient, with point prevalence near or above 1 percent in some Northern regions (Piedmont and Lombardy) but well below .05 percent in Umbria and several Southern regions. These large differences reflect the fact that, while $\Pr(T = 1)$ and $\Pr(T = 0)$ do not vary much, regional variation in $\Pr(P = 1|T = 1)$ is huge.

Figure 3 shows the observed values of $\Pr(T = 1)$ and $\Pr(P = 1|T = 1)$ over time. Both probabilities fall, but the decline in $\Pr(P = 1|T = 1)$ is much stronger. This decline is likely due to, on the one hand, a fall in the intensity of the epidemic and, on the other, a reduction in the bias implicit in the testing criteria. The latter is because the falling number of daily new cases resulting from the containment policies created more room to take swabs from less severe cases and to conduct a more

[9] Following Manski and Molinari (2020) we do not provide measures of statistical precision because we are unsure what type of sampling process would be reasonable to assume for our data.

intense contact tracing, again leading to more swabs taken from asymptomatic subjects.

Since we are not able to quantitatively assess the change in the testing bias, we refrain from showing time changes in our estimates. Rather, we show in Figure 4 how $\Pr(I = 1)$ increases with λ, for a given value of β, on three different dates. This helps to gauge the extent to which a fall in the testing bias, i.e. an increase in λ, would have offset the observed reduction of prevalence among the tested cases ($\Pr(P = 1|T = 1)$), reflected in the clockwise rotation of the three lines. The two horizontal lines show, for example, that prevalence would have remained constant at about 1 percent (.5 percent) if λ had increased from about .15 to 1 (from about .05 to .55).

5 Point vs. period prevalence

As mentioned in the Introduction, we have a similar goal as Manski and Molinari (2020): offering an estimate of the prevalence of the infection based on readily available data and minimal and transparent assumption. The main difference is that we aim to provide estimates of *point* prevalence, while they aim to estimate *period* prevalence. An estimate of period prevalence is in the making in some countries, including Italy, that are organizing a random serological testing of the population. As both concepts of prevalence are interesting and play a different role, we explore here the link between them in order to understand what can be learned about period prevalence from estimates of point prevalence and, more generally, from the data currently available.

5.1 Relationship between the two concepts

Assume for simplicity that the infection lasts at most two periods, nobody dies because of it, and a fraction $1 - \delta$ of the *new* infected recovers after one period while the remaining fraction δ after two periods (the analysis would be qualitatively similar had we considered more complex patterns of recovery). Hence, at any time $t \geq 1$, the number of *currently* infected (i.e., still active carriers of the infection) F_t is

$$F_t = F_t^* + \delta F_{t-1}^*, \tag{9}$$

where F_t^* are the *new* infected at time t, with $F_0^* = 0$ just before the start of the epidemic. The number of people who *just* recovered is

$$V_t = (1 - \delta)F_{t-1}^* + \delta F_{t-2}^*, \tag{10}$$

with $V_1 = 0$ and $V_2 = (1 - \delta)F_1^*$, and the total number of people who are currently recovered from past infections is

$$U_t = \sum_{i=1}^{t} V_i = (1 - \delta)F_{t-1}^* + \sum_{j=2}^{t-1} F_{t-j}^*. \tag{11}$$

Point prevalence at time t in a population of constant size N is

$$\alpha_t = \frac{F_t}{N} = \frac{F_t^* + \delta F_{t-1}^*}{N},$$

while *period* prevalence is

$$\eta_t = \frac{\sum_{i=1}^t F_i^*}{N}.$$

Note that point prevalence is obtained from the *currently* infected, who might have been first infected in previous periods, whereas period prevalence is constructed using the *new* infected in each period. It follows from (9) and (11) that $\sum_{i=1}^t F_i^* = F_t + U_t$. Therefore we have

$$\eta_t = \alpha_t + \frac{U_t}{N}. \tag{12}$$

Equation (12) clarifies the link between the two concepts in the population: starting from point prevalence, to assess period prevalence we need data on the current and past recovered. Equivalently, while point prevalence can be measured using just the data on the currently infected, to estimate period prevalence one also needs the current and past recovered.

We typically do not observe the population, but only a sample drawn from it; the people in the sample are subject to a virological test that reveals – possibly with noise – whether they are currently infected. To further explore the link between the two concepts of prevalence, assume that the sample is a random one, drawn without replacement. Though counterfactual, this provides a tractable and useful benchmark. We maintain the assumption of zero probability of Type-I error and positive probability β of Type-II error.

Thus suppose that, at each time t, a random sample of size n_t is tested for the presence of the virus. If O_t is the number of people who tested positive in the period-t sample, then

$$\mathbb{E}[O_t] = n_t \alpha_t (1 - \beta).$$

Given a reliable estimate of β, we can construct an unbiased estimate of *point* prevalence as $\widehat{\alpha}_t = O_t / [n_t(1 - \beta)]$.

5.2 Data requirement for estimating period prevalence

We could obtain an estimate of *period* prevalence if we also had data recording, in the random samples drawn each period, the number of people recovered from the infection contracted in any of the previous periods. In our simple example, at any date t the fraction of recovered in the population is U_t/N, and therefore the number of recovered in the period-t sample would be (on average) $n_t \, U_t/N$. Dividing by n_t would then give an unbiased estimate of U_t/N and therefore, from (12), an unbiased estimate of period prevalence. Notice however that to identify the recovered, the subjects in the sample should be tested with a *serological* test to detect *past* contacts with the virus.

This is not how the available data are constructed. We know the number of *future* recovered *among the currently infected*, as detected by the *virological* test. But we do not know whether the people who test negative at a point in time *had been infected in the past and are now recovered*, which is precisely the information we would need to compute period prevalence. Hence, even in the ideal case of random sampling, there is a wedge between the data available as by-product of surveillance and treatment protocols and the data that would be needed to construct an accurate estimate of period prevalence. Having a biased sample would only add an extra layer of difficulty, so this conclusion would hold *a fortiori* also in the realistic case of non-random testing.[10]

A wedge between needed and available data is also revealed by the estimate of period prevalence in Manski and Molinari (2020). In keeping with their notation, let C_t be a random variable equal to 1 if, as of day t, a subject has ever been infected and to 0 otherwise, and R_t a random variable equal to 1 if, as of day t, a subject has ever resulted positive to the virological test and to 0 otherwise, i.e. if a subject has never been tested or, if ever tested, resulted negative. With some abuse of notation, for the purposes of this section, also let T_t be a random variable equal to 1 if, as of day t, a subject has ever been tested and to 0 otherwise, i.e. if a subject has never been tested. We can then rewrite equation (10) in their paper, similarly to our equation (6),[11] as

$$\Pr(C_t = 1) = \frac{\Pr(R_t = 1|T_t = 1)}{1 - \beta_t^*} \left[\Pr(T_t = 1) + \lambda_t^* \left(1 - \Pr(T_t = 1)\right)\right], \tag{13}$$

where $\beta_t^* = \Pr(R_t = 0|C_t = 1, T_t = 1)$ and $\lambda_t^* = \Pr(C_t = 1|T_t = 1)/\Pr(C_t = 1|T_t = 0)$. Though superficially similar, β_t^* is not the same as β, the Type-II error probability of a virological test: while β refers to contemporaneous events (i.e., testing negative while infected), β_t^* refers to events that do not necessarily occur at the same point in time.

In general β_t^* is larger than β as the event associated with β_t^* includes, besides all false negatives up to time t, two types of subjects: (i) those tested on day $d < t$ who became infected on a later day $h \leq t$; and (ii) those tested on day $d < t$ who became infected on an earlier day $h < d$ and have fully recovered by day d. The subjects in (i) are not yet infected on the day they are tested, and therefore are correctly recorded as negative on that day, are no longer tested (under the assumption that nobody is tested twice), and yet by day t they have been infected. Similarly, the subjects in (ii) are no longer infected when they are tested, so the virological test correctly finds them as negative, were not tested when infected, and yet by day t they have been infected.

In the simple setting explored in this section, it is easy to verify that, for example,

$$\beta_2^* = \frac{\beta\left[F_1^*(n_1 + \delta n_2) + F_2^* n_2\right] + F_2^* n_1 + F_1^*(1 - \delta)n_2}{(F_1^* + F_2^*)(n_1 + n_2)}, \tag{14}$$

[10] This problem resembles the difficulties faced when trying to estimate the unemployment rate in a population from the stock of currently unemployed people without having data on the flow of new unemployed.

[11] The equivalence between (13) and the estimating equation in Manski and Molinari (2020) is shown in Appendix A.

which is equal to β only if $\beta = \delta = 1$, and is otherwise larger than β for the presence of terms corresponding to (i) and (ii), namely $F_2^* n_1$ and $F_1^*(1-\delta)n_2$ respectively. Appendix B providess the details.

Our key point is that, even in the ideal case of random virological testing, terms of this kind are inherently not observable. While information on β can be obtained from medical expertise and practice, moving from β to β_t^* involves quantities that are not observable with the data currently produced as a by-product of the surveillance and treatment protocols. Therefore, bounds on β_t^* based on bounds on β are unlikely to be tight due to the unobservable (at least with the data readily available) wedge between the two concepts.[12]

6 Conclusions

We showed in this paper how to use the data currently produced as a byproduct of the fight against the pandemic to obtain estimates of the point prevalence of the infection, an important gauge of its evolution.

Our estimates depend on a number of clearly identified features of the pandemic – the Type-I and Type-II error probabilities of the test used to detect the presence of the virus, the relative prevalence of the infection in subsets of the population, the relative size of these subsets in the sample of the tested subjects – about which it is in principle possible to gather background information. The latter, in turn, can be easily incorporated to narrow down the range of the estimate.

We also showed that another important statistic concerning the pandemic – the period prevalence of the infection – is much harder to pin down with the data currently available, and we pointed out which additional information would be needed to obtain such an estimate.

Finally, we applied our method to data from Italy, a country that experienced an early and massive outbreak of the epidemic and is now lifting the social distancing measures introduced to contain the contagion. Our estimates show that in some regions of that country the prevalence of the currently infected might be still large enough to suggest caution in removing all restrictions on mobility, unless an effective system to quickly identify the newly infected, trace their contacts and implement quarantine measures is in place.

We hope that the health authorities, at the national or regional level, by incorporating in our framework their more detailed and disaggregated information about the nature of the pandemic, can obtain a quickly updated gauge of the evolution of the infection, which will help them to contain it.

[12] As explained in Appendix A, Manski and Molinari (2020) impose bounds on $\Pr(C_t = 1|R_t = 0, T_t = 1)$ – which they interpret as (1 minus) the negative predicted value (NPV) of the test – rather than on $\Pr(R_t = 0|C_t = 1, T_t = 1)$. These two probabilities are linked by Bayes Law, and any difficulty in bounding one translates into a difficulty in bounding the other. In fact, the logical possibilities that create a difference between β_t^* and β are also responsible for a difference between $\Pr(C_t = 1|R_t = 0, T_t = 1)$ and (1 minus) the NPV of the test.

References

Flaxman, S., S. Mishra, A. Gandy, H. J. T. Unwin, et al. (2020). Estimating the number of infections and the impact of non- pharmaceutical interventions on COVID-19 in 11 European countries. *Nature*. https://doi.org/10.1038/s41586-020-2405-7.

Istituto Nazionale di Statistica (2019). Resident population as of January 1, 2019. Available at: https://www.istat.it/it/popolazione-e-famiglie/.

Lavezzo, E., E. Franchin, C. Ciavarella, G. Cuomo-Dannenburg, et al. (2020). Suppression of COVID-19 outbreak in the municipality of Vo', Italy. *Nature*. https://doi.org/10.1038/s41586-020-2488-1.

Li, R., S. Pei, B. Chen, Y. Song, et al. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science 386*(6490), 489–493.

Manski, C. F. and F. Molinari (2020). Estimating the COVID-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics*. Forthcoming.

Riccardo, F., M. Ajelli, X. D. Andrianou, A. Bella, et al. (2020). Epidemiological characteristics of COVID-19 cases in Italy and estimates of the reproductive numbers one month into the epidemic. Istituto Superiore di Sanità, Rome. Available at: https://www.medrxiv.org/content/10.1101/2020.04.08.20056861v1.

Stock, J. H. (2020). Data gaps and the policy response to the novel Coronavirus. NBER Working Paper No. 26902.

Vollmer, M. A. C., S. Mishra, H. J. T. Unwin, A. Gandy, et al. (2020). Using mobility to estimate the transmission intensity of COVID-19 in Italy: A subnational analysis with future scenarios. Imperial College COVID-19 Response Team Report 20. Available at: https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-05-04-COVID19-Report-20.pdf.

Watson, J., P. F. Whiting, and J. E. Brush (2020). Interpreting a covid-19 test result. *British Medical Journal* (May 12, 2020). Available at: https://www.bmj.com/content/369/bmj.m1808.

Table 1: Observed values of $\Pr(T = 1)$ and $\Pr(P = 1|T = 1)$, June 20, 2020.

Region	$\Pr(T = 1)$	$\Pr(P = 1\|T = 1)$
Abruzzo	.0005	.0007
Basilicata	.0005	.0000
Calabria	.0004	.0022
Campania	.0001	.0041
Emilia-Romagna	.0008	.0063
Friuli Venezia Giulia	.0008	.0011
Lazio	.0004	.0038
Liguria	.0004	.0106
Lombardia	.0006	.0273
Marche	.0004	.0027
Molise	.0007	.0035
Piemonte	.0004	.0151
Puglia	.0003	.0014
Sardegna	.0005	.0007
Sicilia	.0003	.0008
Toscana	.0004	.0026
Trentino-Alto Adige	.0008	.0069
Umbria	.0005	.0003
Valle d'Aosta	.0006	.0035
Veneto	.0005	.0018
Italy	.0004	.0078

Notes: 7-day centered moving averages of daily changes of tested cases and new detected cases. Source: DPC and Istat.

Table 2: Estimates of $\Pr(I = 1)$ for different values of λ and β, June 20, 2020.

	β										
λ	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
.01	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
.05	.000	.000	.000	.000	.000	.001	.001	.001	.001	.001	.001
.10	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.002
.15	.001	.001	.001	.001	.001	.002	.002	.002	.002	.002	.002
.20	.002	.002	.002	.002	.002	.002	.002	.002	.003	.003	.003
.25	.002	.002	.002	.002	.002	.003	.003	.003	.003	.004	.004
.30	.002	.002	.003	.003	.003	.003	.003	.004	.004	.004	.005
.35	.003	.003	.003	.003	.003	.004	.004	.004	.005	.005	.005
.40	.003	.003	.003	.004	.004	.004	.004	.005	.005	.006	.006
.45	.004	.004	.004	.004	.004	.005	.005	.005	.006	.006	.007
.50	.004	.004	.004	.005	.005	.005	.006	.006	.007	.007	.008
.55	.004	.005	.005	.005	.005	.006	.006	.007	.007	.008	.009
.60	.005	.005	.005	.006	.006	.006	.007	.007	.008	.009	.009
.65	.005	.005	.006	.006	.006	.007	.007	.008	.008	.009	.010
.70	.006	.006	.006	.006	.007	.007	.008	.008	.009	.010	.011
.75	.006	.006	.007	.007	.007	.008	.008	.009	.010	.011	.012
.80	.006	.007	.007	.007	.008	.008	.009	.010	.010	.011	.013
.85	.007	.007	.007	.008	.008	.009	.010	.010	.011	.012	.013
.90	.007	.007	.008	.008	.009	.009	.010	.011	.012	.013	.014
.95	.008	.008	.008	.009	.009	.010	.011	.011	.012	.014	.015
.99	.008	.008	.009	.009	.010	.010	.011	.012	.013	.014	.015

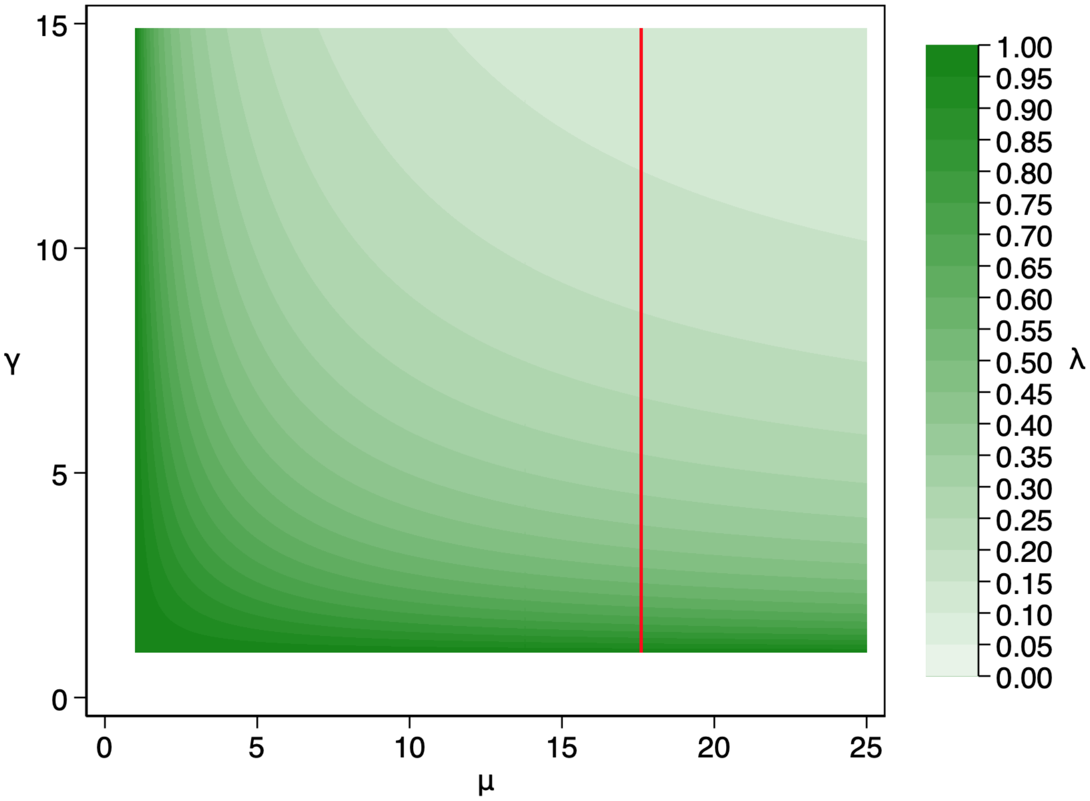
Notes: Values obtained from equation (6) in the main text. The values in red correspond to $.10 \leq \lambda \leq .65$ and $.10 \leq \beta \leq .30$.

Table 3: Estimates of $\Pr(I = 1)$ by region, for different values of (λ, β), June 20, 2020.

	(λ, β)		
Region	$(.25, .10)$	$(.50, .20)$	$(.75, .30)$
Abruzzo	.000	.000	.001
Basilicata	.000	.000	.000
Calabria	.001	.001	.002
Campania	.001	.003	.004
Emilia-Romagna	.002	.004	.007
Friuli Venezia Giulia	.000	.001	.001
Lazio	.001	.002	.004
Liguria	.003	.007	.011
Lombardia	.008	.017	.029
Marche	.001	.002	.003
Molise	.001	.002	.004
Piemonte	.004	.009	.016
Puglia	.000	.001	.001
Sardegna	.000	.000	.001
Sicilia	.000	.000	.001
Toscana	.001	.002	.003
Trentino-Alto Adige	.002	.004	.007
Umbria	.000	.000	.000
Valle d'Aosta	.001	.002	.004
Veneto	.001	.001	.002
Italy	.002	.005	.008

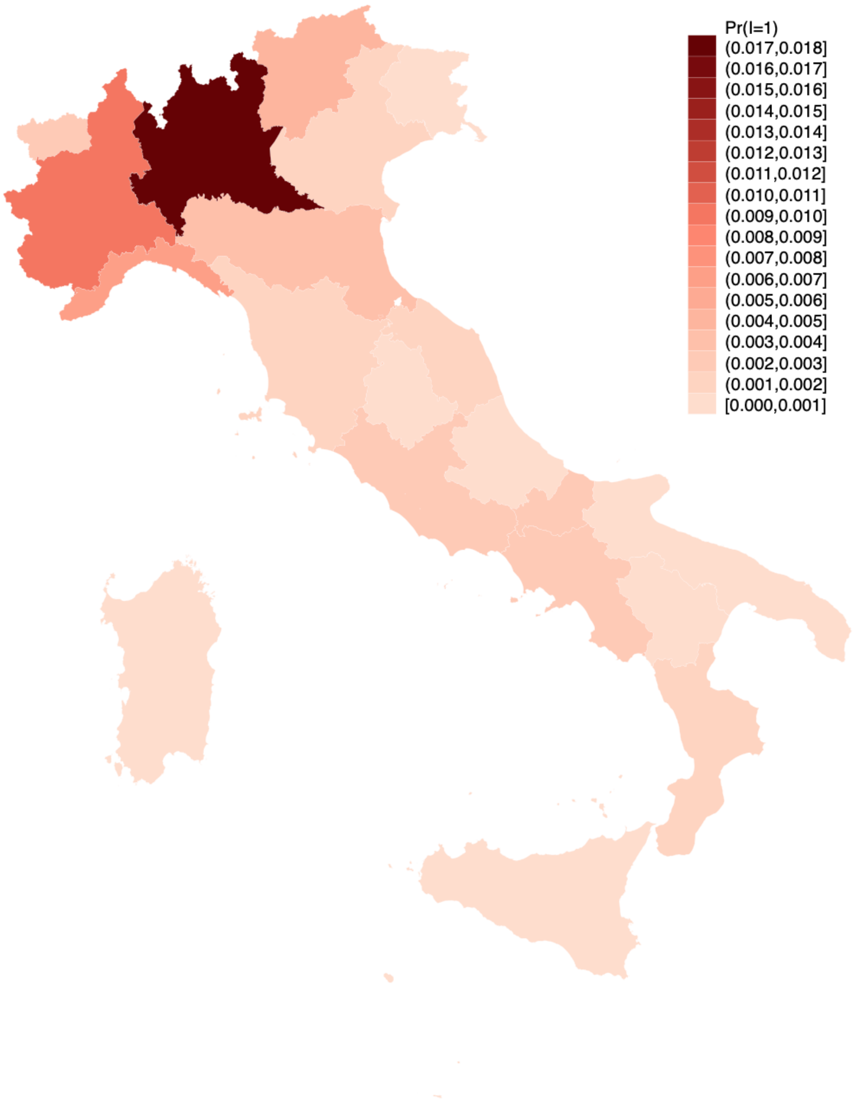
Notes: Values obtained from equation (6) in the main text.

Figure 1: Contour plot of λ as a function of γ and μ for $n = .0004$ and $\pi = .067$.



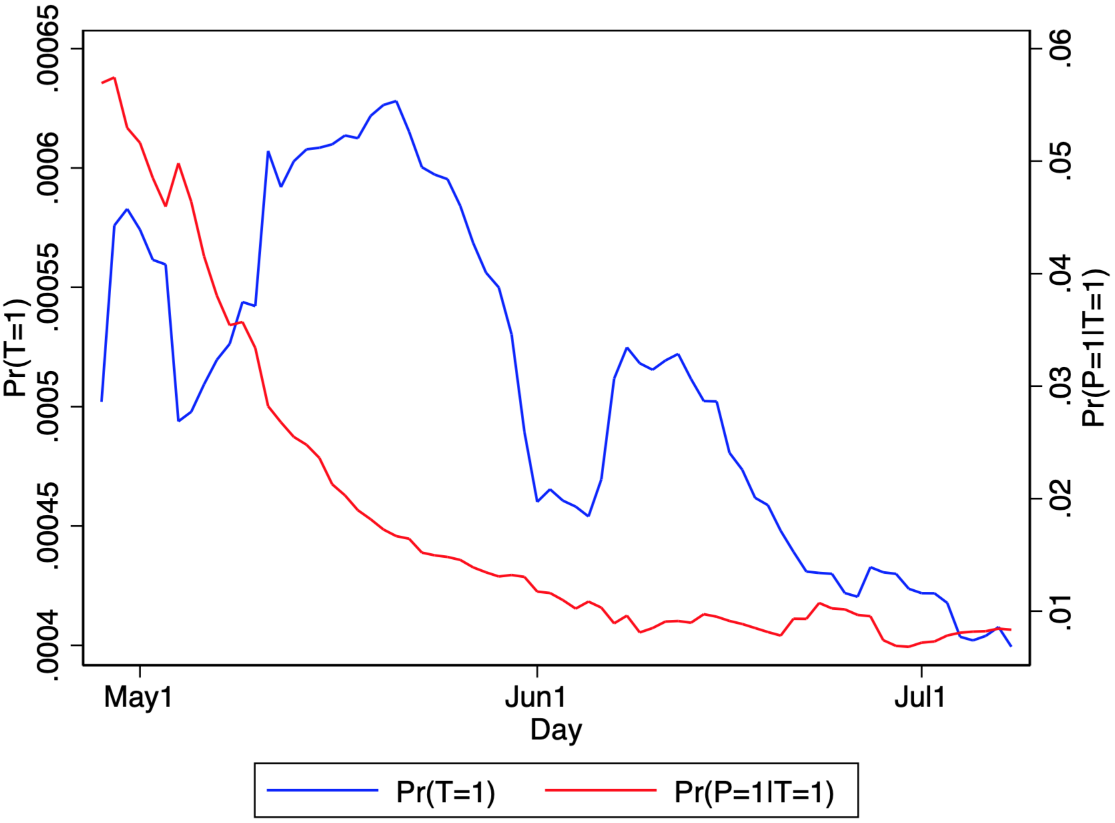
Notes: The contour plot is based on equation (8) in the main text. The red vertical line corresponds to $\mu = 17.9$, the value observed in Vo' on late February.

Figure 2: Estimates of $\Pr(I = 1)$ by region for $\lambda = .50$ and $\beta = .20$, June 20, 2020.



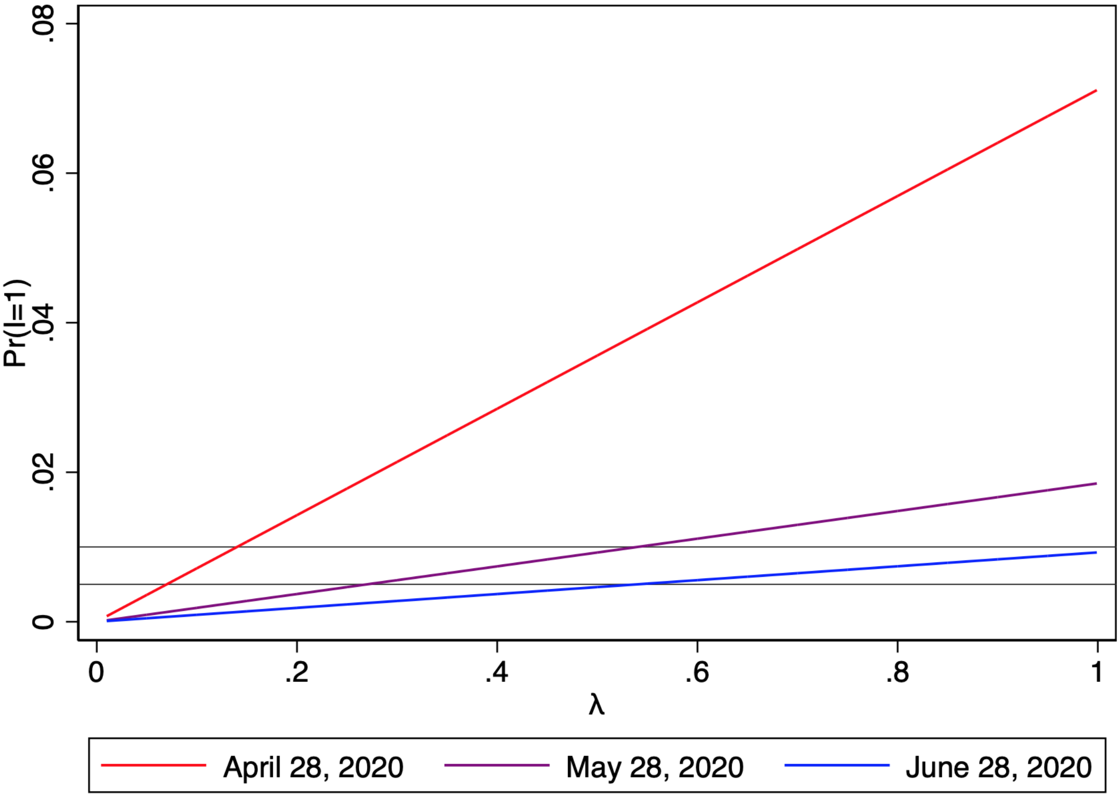
Notes: The figure corresponds to the middle column of Table 3.

Figure 3: Observed values of $\Pr(T = 1)$ and $\Pr(P = 1|T = 1)$ over time.



Notes: 7-day centered moving averages of daily changes of tested cases and new detected cases. Source: DPC and Istat. The scale on the left refers to $\Pr(T = 1)$, the one on the right to $\Pr(P = 1|T = 1)$.

Figure 4: Effect of λ on $\Pr(I = 1)$ for $\beta = .20$.



Notes: The different curves show how the estimate of $\Pr(I = 1)$ would vary as λ varies, for a fixed β, at different dates. The horizontal lines are drawn at .01 and .005.

A Equivalence between two formulations

With the notation introduced in Section 5, equation (10) in Manski and Molinari (2020) can be written

$$\begin{aligned} \Pr(R_t = 1) + B_* \Pr(R_t = 0|T_t = 1) \Pr(T_t = 1) &\leq \\ &\leq \Pr(C_t = 1) \leq \\ &\leq \Pr(R_t = 1|T_t = 1) + B^* \Pr(R_t = 0|T_t = 1), \end{aligned} \tag{15}$$

where B_* and B^* are, respectively, the lower and upper bound on $\Pr(C_t = 1|R_t = 0, T_t = 1)$; the first expression in the chain of inequalities, which provides a lower bound on $\Pr(C_t = 1)$, obtains when $\Pr(C_t = 1|T_t = 0) = 0$, while the last expression, which provides an upper bound on $\Pr(C_t = 1)$, obtains when $\Pr(C_t = 1|T_t = 0) = \Pr(C_t = 1|T_t = 1)$.

We can rewrite (15) as

$$\begin{aligned} \Pr(C_t = 1) = {} & [\Pr(T_t = 1) + \lambda_t^* (1 - \Pr(T_t = 1))] \times \\ & \times [\Pr(R_t = 1|T_t = 1) + \Pr(C_t = 1|R_t = 0, T_t = 1) \Pr(R_t = 0|T_t = 1)] . \end{aligned} \tag{16}$$

Thus $\Pr(C_t = 1)$ is equal to the top term in (15) when $\Pr(C_t = 1|R_t = 0, T_t = 1) = B_*$ and $\lambda_t^* = 0$ – i.e., $\Pr(C_t = 1|T_t = 0) = 0$ – and to the bottom term when $\Pr(C_t = 1|R_t = 0, T_t = 1) = B^*$ and $\lambda_t^* = 1$ – i.e., $\Pr(C_t = 1|T_t = 0) = \Pr(C_t = 1|T_t = 1)$. Now,

$$\begin{aligned} \Pr(C_t = 1|R_t = 0, T_t = 1)\ \Pr(R_t = 0|T_t = 1) &= \frac{\Pr(C_t = 1, R_t = 0, T_t = 1)}{\Pr(T_t = 1)} \\ &= \frac{\Pr(C_t = 1, R_t = 0, T_t = 1)}{\Pr(C_t = 1, T_t = 1)}\ \frac{\Pr(C_t = 1, T_t = 1)}{\Pr(T_t = 1)} \\ &= \Pr(C_t = 1|T_t = 1)\, \beta_t^*, \end{aligned} \tag{17}$$

with β_t^* defined in Section 5. From the definition of λ_t^* and the Law of Total Probability, we get

$$\begin{aligned} \Pr(C_t = 1) &= \Pr(C_t = 1|T_t = 1)\ \Pr(T_t = 1) + \lambda_t^* \Pr(C_t = 1|T_t = 1)\ \Pr(T_t = 0) \\ &= \Pr(C_t = 1|T_t = 1) \left[\Pr(T_t = 1) + \lambda_t^* \Pr(T_t = 0)\right] . \end{aligned} \tag{18}$$

Plugging (17) and (18) into (16), simplifying, and solving out for $\Pr(C_t = 1|T_t = 1)$ yields

$$\Pr(C_t = 1|T_t = 1) = \frac{\Pr(R_t = 1|T_t = 1)}{1 - \beta_t^*}. \tag{19}$$

Substituting (19) into (18) then gives

$$\Pr(C_t = 1) = \frac{\Pr(R_t = 1|T_t = 1)}{1 - \beta_t^*} \left[\Pr(T_t = 1) + \lambda_t^* \Pr(T_t = 0)\right],$$

which is equation (13) in the main text. Note also that, under the assumption of zero Type-I error probability,

$$\Pr(C_t = 1 | R_t = 0, T_t = 1) = \frac{\beta_t^*}{\beta_t^* + \lambda_t^*}. \tag{20}$$

Hence, the bounds on $\Pr(C_t = 1 | R_t = 0, T_t = 1)$ readily translate into bounds on β_t^*.

B Decomposition under random sampling

Assume, as in Section 5, that (i) each person can be infected at most once, the infection lasts at most two periods, nobody dies because of it, and a fraction $1 - \delta$ of the *new* infected recovers after one period while the remaining δ after two periods; (ii) the sample of people tested in each period is random, drawn without replacement; and (iii) the virological test has zero probability of Type-I error and (constant) probability $\beta > 0$ of Type-II error. Using the notation introduced in that section, we can break down the population in each period t into the following groups:[13]

- sampled in period t, currently infected, testing negative: $n_t(F_t^* + \delta F_{t-1}^*)\beta/N$;
- sampled in period t, currently infected, testing positive: $n_t(F_t^* + \delta F_{t-1}^*)(1 - \beta)/N$;
- sampled in period t, newly infected in period $t-1$, currently recovered, testing negative: $n_t(1 - \delta)F_{t-1}^*/N$;
- sampled in period t, newly infected in period $r < t - 1$, currently recovered, testing negative: $n_t F_r^*/N$;
- sampled in period t, not yet infected, testing negative: $n_t(N - \sum_{i=1}^t F_i^*)/N$;
- sampled in period $s < t$, currently infected: $n_s(F_t^* + \delta F_{t-1}^*)/N$;
- sampled in period $s < t$, newly infected in period $t - 1$, currently recovered: $n_s(1 - \delta)F_{t-1}^*/N$;
- sampled in period $s < t$, newly infected in period $r < t - 1$, currently recovered: $n_s F_r^*/N$;
- sampled in period $s < t$, not yet infected: $n_s(N - \sum_{i=1}^t F_i^*)/N$;
- not previously sampled and currently infected: $(N - \sum_{i=1}^t n_i)(F_t^* + \delta F_{t-1}^*)/N$;
- not previously sampled, newly infected in period $t - 1$, currently recovered: $(N - \sum_{i=1}^t n_i)(1 - \delta)F_{t-1}^*/N$;
- not previously sampled, newly infected in period $r < t-1$, currently recovered: $(N - \sum_{i=1}^t n_i)F_r^*/N$;

[13] Some of the items will only make sense for $t > 1$ or $t > 2$.

- not yet sampled, not yet infected: $(N - \sum_{i=1}^{t} n_i)(N - \sum_{i=1}^{t} F_i^*)/N$.

Specializing for simplicity to the case $t = 2$, we can use this decomposition to compute $\beta_2^* = \Pr(R_2 = 0, C_2 = 1, T_2 = 1)/\Pr(C_2 = 1, T_2 = 1)$. On the numerator we must include all the people who never tested positive, were sampled in one of the two periods and were first infected in one of the two periods:

- sampled in period 1, newly infected in period 1, wrongly classified as negative in period 1, negative in period 2 because not tested again: $n_1 F_1^* \beta/N$;
- sampled in period 2, newly infected in period 2, wrongly classified as negative in period 1, negative in period 1 because in that period they were not tested: $n_2 F_2^* \beta/N$;
- sampled in period 2, newly infected in period 1, still infected in period 2, wrongly classified as negative in period 2, negative in period 1 because in that period they were not tested: $n_2 F_1^* \delta\beta/N$;
- sampled in period 2, newly infected in period 1, recovered by period 2, correctly classified as negative in period 2, negative in period 1 because in that period they were not tested: $n_2 F_1^* (1 - \delta)/N$;
- sampled in period 1, newly infected in period 2, correctly classified as negative when tested (in period 1), negative in period 2, because not tested again: $n_1 F_2^*/N$.

As to the denominator, we must include all the people who where first infected, in either period 1 or period 2, and were first tested, in either period 1 or period 2. These are, clearly, $(n_1 + n_2)(F_1^* + F_2^*)/N$. Putting all this together we obtain equation (14) in the main text.